OR-PRM: Process Reward Model for Operations Research

Yilin Wang[1,*] Heng Zhou[1,*] Dongxing Mao[2] Linjie Li[3] Jingru Tan[2] Haochen Han[4,†] Zhengyuan Yang[3] Alex Jinpeng Wang[2,†] Min Li[2]

[1]Zhejiang University [2]Central South University [3]Microsoft [4]Pengcheng Laboratory

*Equal Contribution †Corresponding Authors

Motivation & Problem

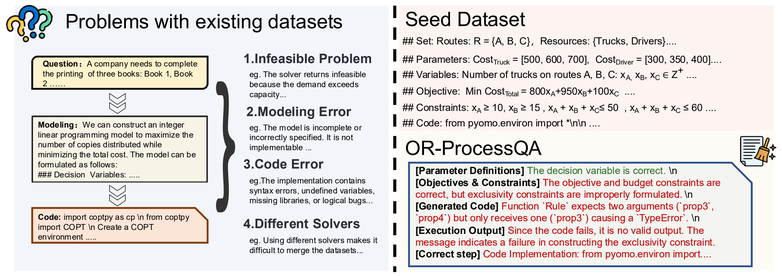


Figure 1. **Left:** Existing noisy datasets. **Right:** Our curated seed data with step-by-step correctness labels.

- OR datasets are **alarmingly unreliable** — over **30%** have serious errors
- Noisy data prevents PRMs from learning faithful reasoning

Key Contributions

- **OR-PRM**: first *Process Reward Model for OR*, providing structured step-level natural language feedback instead of scalar scores
- **OR-ProcessQA**: first OR dataset with reliable step-level correctness labels, built via MCTS + GPT-4o verification
- **+12.5%** average accuracy gain across 6 benchmarks in Best-of-N setting

Case Study: OR-PRM Feedback

OR-PRM Structured Critique Output

[Parameters] The decision variable is correct. ✓

[Objectives & Constraints] Budget constraints are correct, but exclusivity constraints are improperly formulated. ×

[Generated Code] Function `Rule` expects two arguments but only receives one → `TypeError` ×

[Judgement] **Incorrect**

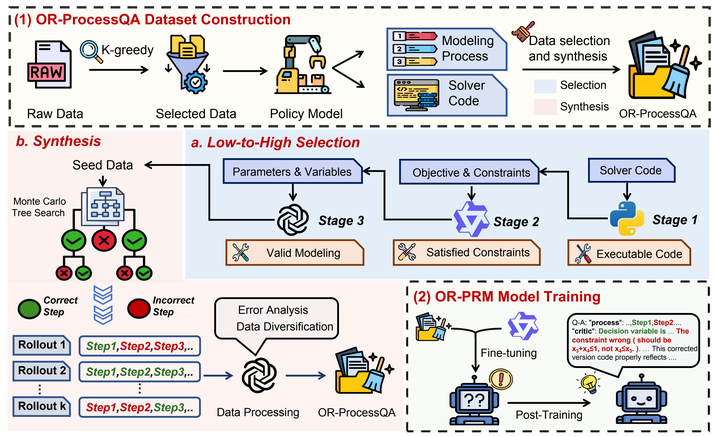
[Corrected Step] The exclusivity constraint should be $x_3 + x_4 \leq 1$, not $x_4 \leq x_3$. Corrected code: `model.addConstr(x3 + x4 <= 1)` ✓

Unlike scalar PRMs, OR-PRM provides **actionable NL feedback** with error localization and corrections.

OR-PRM vs. Traditional PRMs

	Traditional PRM	OR-PRM (Ours)
Output	Scalar score	NL critique
Granularity	Per-step	Per-component
Error type	Binary	Categorized
Correction	None	Auto-fix

Method Overview



Pipeline: (1) Seed dataset via 3-stage filtering. (2) OR-ProcessQA with MCTS + GPT-4o. (3) Train OR-PRM (SFT + DPO).

Data & Training

Seed Data: Three-Stage Validation

- **S1:** Code Execution — error-free solver run
- **S2:** Constraint Check — Qwen3-8B verifies $\hat{x}$
- **S3:** Modeling Accuracy — GPT-4o validates fidelity

OR-ProcessQA: 550K+ Annotated Steps

- MCTS generates trajectories; GPT-4o re-evaluates each step
- Consensus filter: $\text{Label}_{\text{MCTS}} = \text{Label}_{\text{GPT-4o}}$

OR-PRM Training

Stage 1 – SFT: Learn structured NL critique format.

$$\mathcal{L}_{\text{SFT}}(\theta) = -\mathbb{E}_{(x,y)}\left[\sum_{t=1}^{T} \log P_\theta(y_t \mid x, y_{<t})\right]$$

Stage 2 – DPO: Preference alignment (**+8%** over SFT-only).

$$\mathcal{L}_{\text{DPO}} = -\mathbb{E}\left[\log \sigma\left(\beta \log \frac{\pi_\theta(y_w|x)}{\pi_{\text{ref}}(y_w|x)} - \beta \log \frac{\pi_\theta(y_l|x)}{\pi_{\text{ref}}(y_l|x)}\right)\right]$$

Base: Qwen2.5-7B-Coder, 8×A100, DeepSpeed ZeRO-2, bf16.

Conclusion

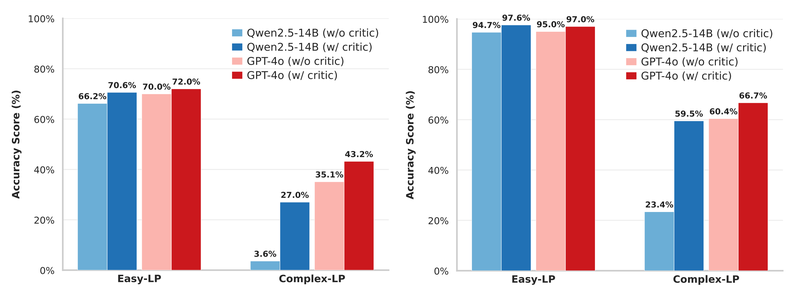
OR-PRM is the **first Process Reward Model for OR**, delivering structured step-level NL feedback. With OR-ProcessQA, it achieves **12.5% avg improvement** (Best-of-N) and **+23.4%** pass@1 as critic, advancing **trustworthy AI reasoning** in OR.

Main Results: Best-of-8

Model	IndOR	Easy	Cplx	NL4LP	NL4O	ReSoc	Avg
GPT-4o	40.5	69.5	35.1	56.2	53.1	47.9	50.4
DeepSeek-v3	66.7	91.9	39.6	92.7	76.5	73.9	73.6
Qwen-7B	19.0	49.7	12.6	50.0	41.3	36.7	34.9
+PRM	23.8	61.8	16.2	56.7	52.1	46.7	**42.9**
							↑+8.0
Qwen-14B	35.7	66.2	3.6	75.8	61.0	50.4	48.8
+PRM	45.2	89.4	12.6	86.5	67.6	66.7	**61.3**
							↑+12.5
Qwen-32B	47.6	80.0	8.2	87.1	68.5	66.3	59.6
+PRM	57.1	96.0	32.4	89.3	74.2	72.7	**70.3**
							↑+10.7
LLMOPT	52.4	96.0	48.6	90.4	81.7	72.2	73.6
+PRM	59.5	97.8	67.6	93.8	85.0	79.2	**80.5**
							↑+6.9

+24.2% on Complex-LP (32B), **+12.5%** avg (14B).

Modeling-Critique-Code Pipeline



(a) pass@1 **(b)** pass@8

Complex-LP: **+23.4%** pass@1 (Qwen-14B), **+8.1%** (GPT-4o).

Ablation Study

Method	Easy-LP	Cplx-LP	Avg
Pass@8 (upper bound)	94.7	23.4	59.1
Self-consistency	50.8	3.6	27.2
Qwen2.5 (Zero-shot)	72.1	9.9	41.0
OR-PRM (SFT only)	79.6	6.3	43.0
OR-PRM (SFT+DPO)	**89.4**	**12.6**	**51.0**

DPO adds **+8.0%** over SFT-only (Qwen2.5-14B).